Package: washoku (via r-universe)

August 31, 2024

Title Extra 'recipes' for Japanese Text, Date and Address Processing

Version 0.0.0.9000

Description In order to handle Japanese text in the feature engineering process, morphological analysis is necessary. Following the framework of `recipes`, to provide steps that can be applied to `textrecipes` for subsequent processing.

License MIT + file LICENSE

Encoding UTF-8

Depends R (>= 2.10), recipes (>= 1.0.0)

Imports generics (>= 0.1.0), magrittr (>= 1.5), purrr (>= 0.3.4), rlang (>= 0.4.8), textrecipes (>= 1.0.0), tibble (>= 3.0.4), vctrs (>= 0.3.4)

Suggests RcppMeCab (>= 0.0.1.2), reticulate (>= 1.17), RMeCab (>= 1.10), sudachir (>= 0.1.0.9000), testthat (>= 2.3.2)

Remotes IshidaMotohiro/RMeCab, uribo/sudachir

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.1

Repository https://uribo.r-universe.dev

RemoteUrl https://github.com/uribo/washoku

RemoteRef HEAD

RemoteSha b10db23d89c88a36afde39e4414e6f185b312890

Contents

step_tokenize_jp *Tokenization of Japanese character variables*

Description

step_tokenize_jp() creates a *specification* of a recipe step that will convert a character predictor into a tokenlist_jp.

Usage

```
step_tokenize_jp(
  recipe,
  ...,
  role = NA,
  trained = FALSE,
  columns = NULL,
  engine = "sudachir",
  options = list(mode = "A", type = "surface", pos = TRUE, instance = NULL),
  skip = FALSE,
  id = rand_id("tokenize_jp")
)
```

Arguments

recipe	A recipe object. The step will be added to the sequence of operations for this recipe.
...	One or more selector functions to choose which variables are affected by the step. See recipes::selections() for more details.
role	Not used by this step since no new variables are created.
trained	A logical to indicate if the quantities for preprocessing have been estimated.
columns	A character string of variable names that will be populated (eventually) by the terms argument. This is NULL until the step is trained by recipes::prep.recipe().
engine	Implement token engine package. Defaults to 'sudachir'.
options	list. path to engine's function.
skip	A logical. Should the step be skipped when the recipe is baked by recipes::bake.recipe()? While all operations are baked when recipes::prep.recipe() is run, some operations may not be able to be conducted on new data (e.g. processing the outcome variable(s)). Care should be taken when using skip = FALSE.
id	A character string that is unique to this step to identify it.

Details

The following packages are available for the engine.

- sudachir (Sudachi)
- RcppMeCab (MeCab)

tokenlist_jp *Token list for Japanese character*

Description

Token list for Japanese character

Usage

```
tokenlist_jp(tokens = list())
```

Arguments

tokens	list

Examples

```
## Not run:
tokens <- list(purrr::set_names(c(intToUtf8(c(26519, 27278))),
               c(intToUtf8(c(21517, 35422)))))
tokenlist_jp(tokens)

## End(Not run)
```

Index